**COURSE : Architectures, systems and algorithms for big data computing**

INSTRUCTOR: Marco Bianchi – Simone Angelini

CONTACTS: +39 06 5480 2123 - +39 06 5480 2809

email : [mbianchi@fub.it](mailto:mbianchi@fub.it) - [sangelini@fub.it](mailto:sangelini@fub.it)

**COURSE BACKGROUND**

This course assumes a basic knowledge of basic data structures (such as lists, hashes and graphs) and a prior knowledge of: SQL, the Python programming language and the linux shell commands.

**LEARNING OBJECTIVES**

By the end of this course, students will know:

* principles of MapReduce programming paradigm, reinforced by presentation of some algorithms implemented in MapReduce style;
* main features and architectural components of the following big data frameworks: Hadoop, Storm, Hive, Spark;
* how to analyse and select the proper technologies to face a Big Data problem;
* the “Lambda-Architecture” pattern;
* how to learn by-example using a pre-configured Big Data Platform.

**METHODOLOGY**

Lecture/presentation, discussion, question and answers, demonstrations, practical sessions (hands-on practice).

**EXAM**

Small project and oral exam.

**CONTENTS**

Introduction to big data problems and platforms. Map-Reduce and examples: word count, average temperature, image smoothing, page-rank. Hadoop, HDFS. Hadoop2 in practice: data logistics (data serialization, organizing and optimizing data in HDFS, moving data into and out of Hadoop), big data patterns (joining), data structures and algorithms at scale (e.g. Bloom Filters). Beyond MapReduce: SQL on Hadoop (Hive). Outline of Apache Storm and Apache Spark. Lambda architecture. Laboratory sessions focused on Hadoop, Hive and Spark.

**TEACHING MATERIAL**

Slides and references at free resources on the Web. Some of these are:

* [MapReduce: Simplified Data Processing on Large Clusters](http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf)
* [The Google File System](http://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf)
* [Hive – A Petabyte Scale Data Warehouse Using Hadoop](http://infolab.stanford.edu/~ragho/hive-icde2010.pdf)
* Jure Leskovec, Anand Rajaraman, Jeff Ullman - *Mining of Massive Datasets* – free ebook: http://www.mmds.org

**SUGGESTED READING**

* Alex Holmes - *Hadoop on Practice (second edition)* - 2015 – Manning
* Allen, Jankowsky, Pathirana – *Storm Applied – strategies for real-time event processing* – 2015 Manning
* Nathan Marz – *Big Data – principles and best practice of scalable real-time data system* – 2015 – Manning
* H. Karau et al. - Learning Spark: Lightning-Fast Big Data Analysis – 2015 – O’Reilly

**ADDITIONAL SUGGESTED TEXTBOOKS**

* Grover, Malaska, Seidman & Shapira - *Hadoop Application Architectures Designing real-world big data application – 2015 – O’Reilly*
* S. Ryza et al. - Advanced Analytics with Spark: Patterns for Learning from Data at Scale - *2015 – O’Reilly*